Uniqueness and How it Impacts Privacy in Health-Related Social Science Datasets

A. Cheyenne Solomon
School of Informatics and Computing
Indiana University
Bloomington, Indiana
aychsolo@cs.indiana.edu

Raquel Hill
School of Informatics and Computing
Indiana University
Bloomington, Indiana
ralhill@cs.indiana.edu

Erick Janssen
Kinsey Institute for Research in Sex, Gender, and Reproduction
Bloomington, Indiana
ejanssen@indiana.edu

Stephanie A. Sanders
Kinsey Institute for Research in Sex, Gender, and Reproduction
Bloomington, Indiana
sanders@indiana.edu

Julia R. Heiman
Kinsey Institute for Research in Sex, Gender, and Reproduction
Bloomington, Indiana
jheiman@indiana.edu

ABSTRACT

Social scientists, like those performing research at the Kinsey Institute for Research in Sex, Gender and Reproduction, may use surveys to gather large amounts of sensitive data. Unlike purely medical-related datasets, these social science datasets tend to be sparse and high-dimensional, which presents opportunities to characterize participants in the dataset in unique ways. These unique characterizations may enable individuals to be linked to external data in ways that have not been previously considered. Therefore, traditional approaches to de-identifying data, such as fulfilling HIPAA requirements, may not be sufficient for preventing the re-identification of participants in large social science datasets.

In this paper, we evaluate the statistical characteristics of two high-dimensional social science datasets to better understand how unique features impact privacy. We apply a class of statistical de-anonymization attacks in an attempt to achieve theoretical re-identification of participants. We assume that an attacker has exact knowledge of a subset of attribute values for a particular record, and wants to link this subset of data to the actual record to discover the remaining content. We show that although 98% of the records within the dataset are unique given any three attributes, re-identification of the records may not be easily achieved. We attribute limited re-identification to the inherent similarity in the human behavior that the scientists measure. This work is the first to characterize re-identification risks in high-dimensional data that is collected in surveys designed to capture the various behaviors and experiences of groups of individuals.

IHI'12, January 28–30, 2012, Miami, Florida, USA.

Categories and Subject Descriptors

K.4.1 [**Public Policy Issues**]: Data Privacy; J.3 [**Life and Medical Sciences**]: Medical information systems; J.4 [**Social and Behavioral Sciences**]: Sex research

General Terms

Security

Keywords

Privacy, re-identification, uniqueness, similarity

1. INTRODUCTION

The sharing of medical data has many advantages, including: the creation of a unified data display for clinicians, the development of predictive and diagnostic support systems, reductions in institutional costs, and improvements in medical care. Medical data are often not shared with external parties because even when data is de-identified per HIPAA Safe Harbor rules, sharing of such data may introduce privacy risks. This is because privacy does not only depend on de-identified data but also on context-specific information, such as shared demographic data of subjects, presence of fields that may be linked across other existing databases, social relationships among subjects, and profiles of the data recipient.

Preserving the privacy of medical-related data is even more challenging when the data are obtained from studies that create unique profiles of individual participants. These participant profiles can be so unique that traditional anonymization techniques cannot be used to generalize and de-identify the record. Therefore sharing these data with external parties may become a lengthy process of negotiating specific use agreements. Sharing of the data among researchers within the organization that owns the data also risks privacy. Even when traditional identifiers are removed, the uniqueness of these records may make re-identification easy for anyone who has access to the complete record.

Privacy is important to social science researchers, like those at The Kinsey Institute for Research in Sex, Gender, and

Reproduction, because they depend upon participants providing accurate answers about their personal behaviors. In this paper we evaluate two datasets from the Kinsey Institute. These datasets were specifically collected as part of research projects designed to enhance the understanding of behavioral and psychosocial factors related to risk for human immunodeficiency virus (HIV) and other sexually transmitted infections (STI) as well as other risks to well-being. As sexuality-related data are considered sensitive, they require protection of both confidentiality and privacy—a shared feature with other types of health data. If participants provide inaccurate information, researchers may make incorrect conclusions with consequences for public health. The protection of data like these serves two purposes: to protect participants' privacy and, through that, to help encourage them to give accurate answers to increase the quality of research. For participants to give accurate information, they must trust that the researchers will strive to protect their information from breaches in confidentiality and invasions of privacy. Researchers must balance three interests in the development and sharing of datasets: 1) the desire to collect data on a range of variables that alone or in combination may risk re-identification; 2) the increasing pressure from the scientific and medical communities as well as funding and regulatory agencies that datasets be made available to others; and 3) human subjects concerns such as potential breaches in confidentiality and privacy.

The goal of this work is to create an understanding of uniqueness and re-identification vulnerability in high-dimensional social science datasets that contain demographic, medical and behavioral data. This data is high-dimensional because each record contains many attributes and each attribute may have many possible values. Each attribute/value combination can be viewed as a dimension. Such high-dimensional microdata (i.e. information about specific individuals) are often sparse, meaning that there are no similar records within the multi-dimensional space.

We make the following contributions: First we analyze the data and characterize uniqueness and sparsity of records within the dataset. We also introduce a uniqueness similarity measure which helps to further characterize the data. The results of this analysis are not limited to the datasets that we evaluated, but may be applied to datasets that have similar size and are constructed using similar survey techniques. Additionally, we adapt a class of de-anonymization attacks by Narayanan and Shmatikov [7] and demonstrate the likelihood of re-identification when an attacker has exact information. We use exact information to establish the best-case scenario for an adversary trying to execute re-identification. We introduce error to the auxiliary information to create a more realistic attack scenario. We assume that an attacker has some subset of attribute values for a specific record and some information that identifies the individual associated with the record. The attacker's goal is to link the partial information to the complete record. We limit the amount of information that an attacker may have, but not the type.

The outline of this paper is as follows: Section 2 is related work in re-identification, Section 3 details our methods and results, Section 4 presents our discussion and interpretation of those results, and Section 5 concludes with future and ongoing work.

2. RELATED WORK

Medical datasets contain information that both patients and providers have a vested interest in keeping private. Open government laws and data-sharing policies of funding agencies, such as the National Institutes of Health and the National Science Foundation, may require these data to be shared in some form. This sharing may pose substantial risk to privacy.

As we have seen from Sweeney's work with Massachusetts hospital discharge records [9], there is a risk of re-identification even from sanitized datasets. Sweeney was able to re-identify the then-governor of Massachusetts by linking the hospital discharge record to a voter registration database. She used only his birth date, gender, and zip code. This work introduced the concept of looking not only at what pieces of information are *explicitly* identifying, such as name, social security number, address, and phone number, but also what *could be* identifying. Such variables are considered quasi-identifiers, information with which an adversary can re-identify a person by linking it with other datasets.

Genomic data is a type of medical microdata receiving extensive attention. Malin and Sweeney linked many genomic datasets together to re-identify people in [5] using disease genes. The presence of a disease gene in a medical facility's record indicates a patient visit, and these visits were correlated to re-identify. In similar work [6], they linked disease genes with anonymized clinical and DNA reports from hospitals. Malin identified family relations in [4] by linking de-identified pedigrees to online genealogy datasets and online newspaper death records. De-identified pedigrees consist of only family structures with gender and death status. Notable here is that the author did not use official data sources (hospital, governmental, etc.). Instead, he used possibly erroneous online sources and validated his findings with the Social Security Death Index. This is an example of how unofficial sources can just as easily lead to privacy risks.

Some of those unofficial sources have been considered impervious. In 2006 AOL released three months of anonymized Internet searches. This led to the re-identification of some users using only their search queries [1]. Likewise, social networks' privacy weaknesses are becoming more and more of a concern, even when it is possible to make profiles 'private' [11]. Narayanan and Shmatikov have evaluated the Netflix Prize dataset [7], which had been 'anonymized' by removing all information other than movie ratings and date of rating, for each movie. While this information is not linkable to any outside governmental database, an adversary who has some prior knowledge of an individual's taste in movies can reconstruct an entire Netflix Prize record.

While there has been extensive work in medical-related datasets, very large and sparse microdatasets, and social networks, no previous work looks at how social science datasets can pose a risk to participant privacy. They are unlike medical-related and microdatasets because individual attributes are much richer. They are also far more high-dimensional than medical-related datasets tend to be and less sparse than microdatasets. This work builds upon our previous work in [8].

3. RESULTS

We explore uniqueness, similarity, and re-identification for two datasets from The Kinsey Institute. The datasets con-

tain participants who have answered some subset of 332 questions from a sexual health survey. The survey consists of multiple modules, including: Background, Mood and Sexuality Questionnaire (MSQ), Sexual Inhibition and Sexual Excitation Scales (SIS/SES), Kinsey Institute Sexual Behavior Questionnaire (KISBQ), State-Trait Anxiety Inventory (STAI) and Zemore Depression Proneness Rating (ZDPR). MSQ assesses how mood (i.e. anger/frustration, anxiety/sadness, happiness/cheerfulness, and sadness/depression) affects sexuality. SIS/SES measures sexual inhibition and sexual excitation scales. KISBQ evaluates various risky sexual behaviors. STAI measures anxiety proneness. ZPDR estimates depression proneness. Questions from SIS/SES, MSQ, STAI, and ZDPR are largely multiple choice and those within background and KISBQ are a mix of multiple choice, numerical entry, and text entry. Important to note is that SIS/SES, MSQ, STAI, and ZDPR are standardized sexual health scales. A scale is a questionnaire with which you hope to measure a psychological construct, with multiple questions (also called items or, for scales, indicators) [2]. In general, the steps for developing a scale include the development of questions that are relevant to the construct. After a large number of subjects answer all the questions, factor analysis is used to select items that are relevant to the construct and eliminate ones that are not. Factor analysis removes items that are very skewed (e.g., if 90% answer that they 'can be shy at times', that is not a good item to measure introversion, as you hope to be measuring a trait in which people vary). For our Kinsey datasets, the scores from the scaled modules SIS/SES, ZDPR, and STAI approximate a normal distribution. Scores from MSQ do not generally follow a normal distribution.

Dataset 1 contains 10865 participants who were recruited from a small Midwestern town using newspaper advertisements, fliers, etc. Dataset 2 contains 5931 participants and is a convenience sample of individuals within the USA that visited the Kinsey Institute's website. The background and demographic data for participants in Dataset 1 are more similar than for Dataset 2, which allows us to compare and contrast how uniqueness and similarity impact re-identification. All participants within both datasets have answered some subset of 332 sexual health survey questions. Participants from Dataset 1 were given a 20-question experimental module. This module was not included in the survey that was administered to participants in Dataset 2. Dataset 1 was not processed for missing values, nor were extreme values recoded or outliers removed. Dataset 2 has been processed for missing values.

3.1 Uniqueness

As Malin states in [4], one needs uniqueness and linkage to successfully re-identify. Our uniqueness analyses suggests that almost any attribute, either by itself or in combination with another, can be used to make a participant unique. Even answers to multiple choice questions, such as whether one has cheated on one's spouse (yes, no, or not applicable), can be combined to uniquify[1] participants.

[1]The term 'uniquify' refers to the reduction of the query population to a single participant. A participant's answers to a survey question or set of survey questions is unique if they are the only person to provide those specific answers.

Evaluation Category	% Total	% NT*
Uniquified by 1 answer	36.31	3.76
Uniquified by 2-answer combinations	79.27	54.61
Uniquified by 2-answer combinations, but not 1 answer	42.96	–
Uniquified by 3-answer combinations	97.6	–
Uniquified in more than 100 ways	14.88	7.48
Uniquified in more than 20 ways	36.52	22.77

* NT: No text-entry answers. Text-entry answers are most often trivially unique, e.g. "Aderall, Advaire" could be unique due to misspellings. We believe that omitting them gives a more accurate picture of uniquification.

Table 1: Initial results of simple data analyses on Dataset 1.

Evaluation Category	% Total	% NT
Uniquified by 1 answer	52.87	5.56
Uniquified by 2-answer combinations	83.70	68.61
Uniquified by 2-answer combinations, but not 1 answer	37.31	–
Uniquified by 3-answer combinations	100	–
Uniquified in more than 100 ways	23.92	21.02
Uniquified in more than 20 ways	50.59	44.82

Table 2: Initial results of simple data analyses on Dataset 2.

We ask the following questions of the data:

- How many participants are unique in various combinations of fields?
- How many fields must be combined before 100%, or nearly 100%, of participants have at least 1 unique answer or combination of answers?
- Which fields make participants more vulnerable than others?
- Are text-entry fields more vulnerable than multiple-choice fields?
- Does partitioning the fields, such as by survey module, affect a participant's vulnerability?
- Do participants provide unique answers to the same questions?[2]
- Can uniquification predict other factors about a participant?[3]

Tables 1 and 2 summarize results and show that even when considering only two data fields, or survey questions, over 50% of the participants within the datasets are unique. When

[2]This question is explored in the following section in Uniqueness Similarity.

[3]This question is more relevant to social sciences, as it refers to behavioral predictions, but has implications for privacy as well.

Module	% Singles	% Doubles
Background	26.96	36.68
Demographics	3.40	13.57
General Health	23.77	5.41
Sexual Behavior	1.84	7.72
MSQ	.03	.62
KISBQ	2.62	6.71
SIS/SES	.01	.01
STAI	.01	.02
ZDPR	.12	.81

Table 3: Results of intra-module uniqueness, looking only within the stated module.

Module	% Singles	% Doubles
Background	39.56	49.06
Demographics	4.84	15.39
General Health	35.69	8.34
Sexual Behavior	3.33	13.64
MSQ	0	1.99
KISBQ	7.23	13.79
SIS/SES	0	.73
STAI	0	.017
ZDPR	0	2.43

Table 4: Results of intra-module uniqueness for Dataset 2, as in Table 3.

considering three attributes, everyone in Dataset 2 and almost everyone in Dataset 1 has at least one unique combination. To determine uniqueness, we compare each participants' answers to every other participants' answers. Combinations are unique if the combination of both answers is unique, not if both in the combination are singly unique. Answers for which a participant is singly unique are not considered in combination analysis because all combinations with such answers would be trivially unique. In combinations of three, combinations of two that would make participants trivially unique in many combinations of three were not removed, but singly unique attributes were. Dataset 2 shows higher uniquification rates in all categories. This is probably due to population size, though population origin (around the US as opposed to only within a single Midwestern town) may also have contributed.

Tables 3 and 4 summarize results in intra-module uniqueness analysis. The modules are split into scaled modules and unscaled. The results show higher uniquification rates for Background and KISBQ modules for Dataset 2 in comparison to Dataset 1. We attribute this increase to the more diverse population in Dataset 2. Additionally, individuals with similar backgrounds also exhibit similar risky sexual behavior (i.e. KISBQ). Most of the modules are standardized sexual health scales, which may produce a normal distribution of responses. Given a normal distribution, we expect participant answers to be less unique, which explains why so few participants are uniquified within scaled modules.

Evaluation Category	# Questions
Uniquifies no one, singletons:	230
Uniquifies no one, 2-combinations:	29

Table 5: This table shows how many questions are not used in uniquification of any participant.

Evaluation Category	# Questions
Uniquifies no one, singletons:	259
Uniquifies no one, 2-combinations:	2

Table 6: As in Table 5, but for Dataset 2. Dataset 2 contains 312 fields to Dataset 1's 332.

Tables 5 and 6 present the number of questions for which there are no unique answers. When considered individually, most questions are not uniquifying, but almost 90% of questions present can be used in combination with some other question to uniquify at least one participant (or 99% for Dataset 2).

Results in uniqueness suggest that correlating answers to multiple questions may create significant re-identification vulnerabilities. If all participants have at least one unique combination of attributes starting with as few as three, the adversary's challenge becomes finding those combinations.

3.2 Similarity Measures

In addition to what makes participants unique, we have investigated ways by which participants are similar. We use a modified cosine similarity as used in [7] for our general similarity:

$$\frac{\Sigma\ Sim(r1_i, r2_i)}{||support(r1) \cup support(r2)||} \quad (1)$$

The support is the set of attributes for which a participant has a numerical value. Sim outputs 1 if the value for attribute i for both records, $r1$ and $r2$, is the same and 0 otherwise. Our results for the nearest neighbor over all attributes for Datasets 1 and 2 are found in Figure 1. We see that participants' nearest neighbors tend to be very close compared to participants in the Netflix Prize dataset[7].

Indeed, where no Netflix Prize participants had a nearest neighbor over 50% similar, over 75% of the participants in Dataset 1 are at least 50% similar to their nearest neighbor. Dataset 2 exhibits a sharp decline in similarity, as only 2.5% are at least 50% similar but over 84% are at least 40% similar. This suggests that the datasets are far less sparse than the Netflix Prize dataset.

Figure 2 examines how similar participants are with modules. In Dataset 1 not every participant has answered every module, which is why there is a large percentage with a nearest neighbor at 0% for the KISBQ and STAI modules. In Dataset 2, nearly every participant has answered every module (though not every question within every module). These graphs give an idea of where the similarity most likely comes from in overall nearest neighbor. Our similarity results confirm that the background of participants in Dataset 1 are more similar than in Dataset 2. Similarly, while risky sexual behaviors (KISBQ) has high similarity results for both

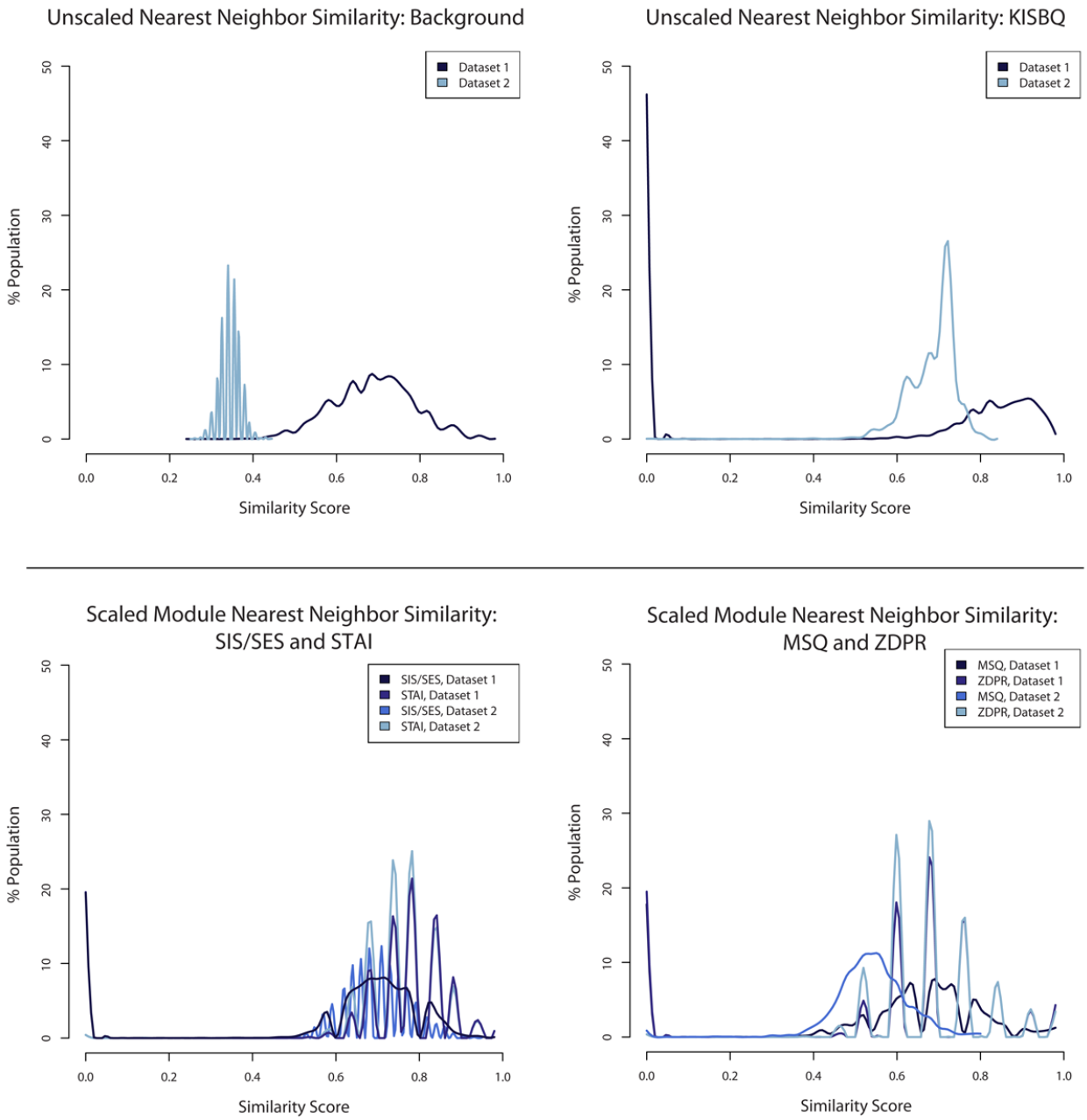

Figure 2: Line histograms of similarity to nearest neighbor within modules. The scaled modules (MSQ, STAI, ZDPR, and SIS) have predictably higher similarity trends than the unscaled surveys (background and KISBQ). Dataset 1 has a significant proportion of 0% similar neighbors in modules because not every participant took every survey. Dataset 2 has been cleaned to contain only those that answered every survey.

datasets, the similarity values for Dataset 1 are much higher. We attribute this increased similarity to the highly similar backgrounds for participants in Dataset 1. Additionally, for scaled modules, we have consistently higher similarity values for the two datasets. Given a normal distribution, we expect participant answers to be more similar, which explains why the two datasets have consistent similarity results for scaled modules.

The second similarity measure is *uniqueness similarity*. It shows how participants' uniquification can be similar (e.g. many participants have unique values for number of sexual partners, but very few have unique values for multiple choice questions within scaled modules).

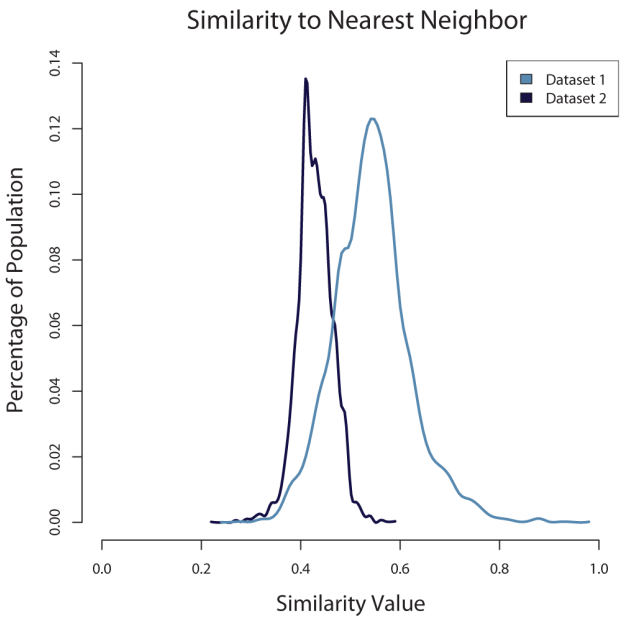

Figure 1: Line histograms of similarity to nearest neighbor for Datasets 1 and 2. Unlike the Netflix dataset on which our similarity work is based, the nearest neighbor in the Kinsey datasets tend to be fairly similar.

$$\frac{\Sigma\ Unique\_Value(r1_i, r2_i)}{||usupport(r1) \cup usupport(r2)||} \quad (2)$$

Here, $usupport$ is the size of the set of unique features. The $Unique\_Value$ function outputs 1 if $r1$ and $r2$ have (a) unique value(s) for attribute set i and 0 otherwise.

Figures 3 and 4 give our results for uniqueness similarity in singletons and doubles. In singletons, most participants have no unique attributes and participants that do tend to have only one. When participants are unique, they have unique values for the same attributes. In combinations of two attributes, most participants are unique in at least one combination. Those that are unique again tend to be unique in the same attributes, thus the spike at 100% similarity.

3.3 Experimental Re-Identification

In previous sections, our results show that most participants have unique characteristics, which may lead one to assume that they may be easily linked to some external, identifying information. On the other hand, our results also show that participants have higher similarity values than those in other microdata. Therefore, in this section we perform experimental re-identification to assess how uniqueness and similarity impact an attacker's ability to re-identify.

We have adapted Narayanan and Shmatikov's simple `scoreboard` algorithm [7] to perform a theoretical re-identification of participants in the Kinsey Institute datasets. The scoreboard algorithm works as follows:

- Select a participant record and extract a random selection of attributes from it as the *auxiliary information.*
- Compare the auxiliary information against all of the records in the dataset using `LesserSim` (essentially the numerator of Equation (1)):

  `LesserSim`: *output* 1 *if all attributes match,* 0 *otherwise*

  If any attribute does not match, `LesserSim` fails (the record is not similar to the auxiliary information).
- If a record matches according to `LesserSim`, add it to the matching set. If only one record is in the matching set after the algorithm terminates and no error has been added, the record is re-identified. If error has been added, it could be a false positive.

Our theoretical adversary has some subset of attribute values for a specific record. The attacker's goal is to link the partial

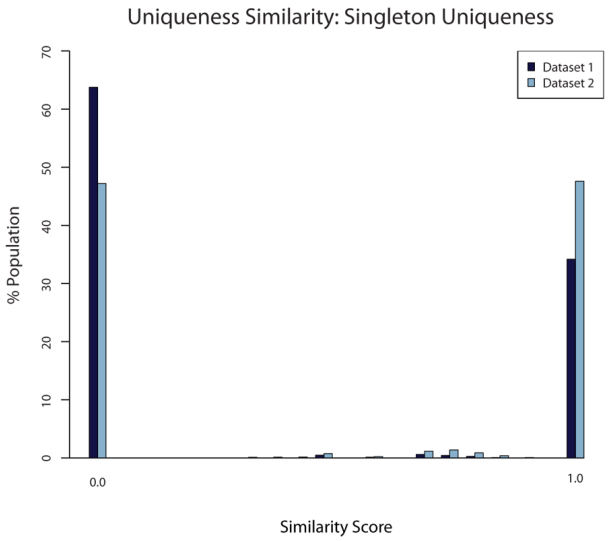

Figure 3: Histogram of nearest neighbor uniqueness similarity for singletons. This graph depicts how often participants are unique *in the same attributes.*

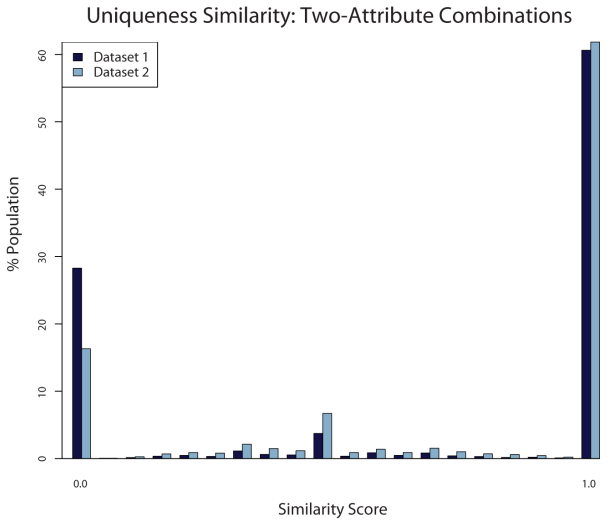

Figure 4: As in Figure 3, but for combinations of two attributes.

information to the complete record. We limit the adversary within the following experimental parameters:

- the number of attributes in the auxiliary set,
- the subset of attributes from which the auxiliary information is taken, either randomly chosen or specifically chosen, and
- whether random error is applied to the auxiliary information.

The amount of error that can be applied to a given attribute depends on the range of possible values for that attribute. Numerical entry attributes and scaled scores have a far larger range of permissible values than multiple choice and this is reflected in error that is added to them. When calculating similarity, we extend the threshold for acceptable values to reflect the amount of error that can be added.

We perform two sets of experimental re-identifications. In the first set of experiments, we randomly choose auxiliary information from either the entire attribute set or from specific modules. In the second set of experiments, the auxiliary information consists of the most uniquifying attributes.

In the first set of experiments, we performed 500 experiments on each participant for randomly selected attributes. A subset of four or nine randomly chosen attributes is formed from a participant's record and matched against the dataset to see if we can reduce the re-identification set to 1 (the correct match). We took the median of the size of the reported matching set for each participant's set of trials. Figures 5 and 6 present the center and spread[4] of the number of records that match the auxiliary information. If there is only one record that matches the auxiliary information, the experiment is 'successful': the adversary has re-identified the target. In cases where error has been added to the auxiliary information, this could be a false positive—instead of finding the record that sourced the auxiliary information, it instead finds an entirely different record.

Tables 7 and 8 show the frequency of re-identifications when we select attributes based on which are the most uniquifying. One set of experiments excludes scaled scores. Although they are amongst the top uniquifying attributes, it is likely that only insiders would know a participant's scores. Calculated scores are not shared with participants. To simulate a case where the adversary is not an insider and could conceivably know anything that *the participant also knows*, we remove calculated scores from auxiliary information. We randomly chose a set of attributes to act as controls. They were chosen once and used for both datasets. These attributes have a mix of uniquification rates and represent a typical case in that we did not select them for their uniquifying properties. In all experiments with added error, we ran 20 trials for each participant. Without added error, we run only one experiment per individual because there are no random elements.

When the adversary has exact knowledge and the size of the matching set is one, a participant is re-identified. When error is added, experiments can result in false positives and misses. False positives occur when the size of the matching set is one, but the record is not the same as the source of the auxiliary information. Misses occur when the size of the matching set is zero, that is, when adding error has brought the auxiliary information out of range of all records. Table 8 gives results for selected attributes with added error. It gives rates of re-identifications, false positives, and misses.

[4]The median is the thick central line in the boxes. The upper and lower horizontal lines signify the third and first quartile respectively. The small lines attached to the vertical lines (the 'whiskers') represent the minimum and maximum. Circles beyond the minima and maxima are outliers, calculated as being outside $3 \cdot median$.

	% Re-Identifications	
Attr. Set	Dataset 1	Dataset 2
Top 9	0	100
Top 9, No Scores	0	99.56
Random 9	0	42.20
Top 4	0	97.47
Top 4, No Scores	0	93.24
Random 4	0	1.92

Table 7: Percent of participants re-identified when we specifically select certain attributes for the adversary's auxiliary information. Here, the auxiliary information is correct.

	% Re-identifications	
Attr. Set	Dataset 1 (*)(**)	Dataset 2 (*)(**)
Top 9	0 (0)(0)	3.59 (31.99)(11.24)
Top 9, No Scores	0 (0)(0.03)	4.47 (4.91)(3.11)
Random 9	0 (0)(37.46)	0 (60.59)(10.70)
Top 4	0 (0)(0)	1.03 (2.46)(2.09)
Top 4, No Scores	0 (0(0)	0.51 (0)(0)
Random 4	0 (0)(2.18)	0.01 (9.48)(1.52)

*: % trials resulting in matching set of size 0

**: % false positive

Table 8: As in Table 7, but error was added to the auxiliary information. False positive and miss rates are also given.

4. DISCUSSION AND CONCLUSIONS

We have shown that uniquification rates for our datasets are notably high. While few individuals have singly unique attributes, combining attributes (e.g. 'Who over the age of 30 has had 10 or more sexual partners in the last year?') yields very high uniquification rates. When considering combinations of three attributes, the percentage of unique records approaches or meets 100%. Our evaluation of Dataset 1 shows that participants with similar backgrounds provide similar answers to questions about sexual behavior and psychological state. Despite the high uniquification rates, records tend to be very similar to their nearest neighbors. In Dataset 1, more than 75% of records have nearest neighbors that are at least 50% similar across all attributes. Participants in Dataset 2 have dissimilar backgrounds and therefore the answers that they provide are predictably more unique and less

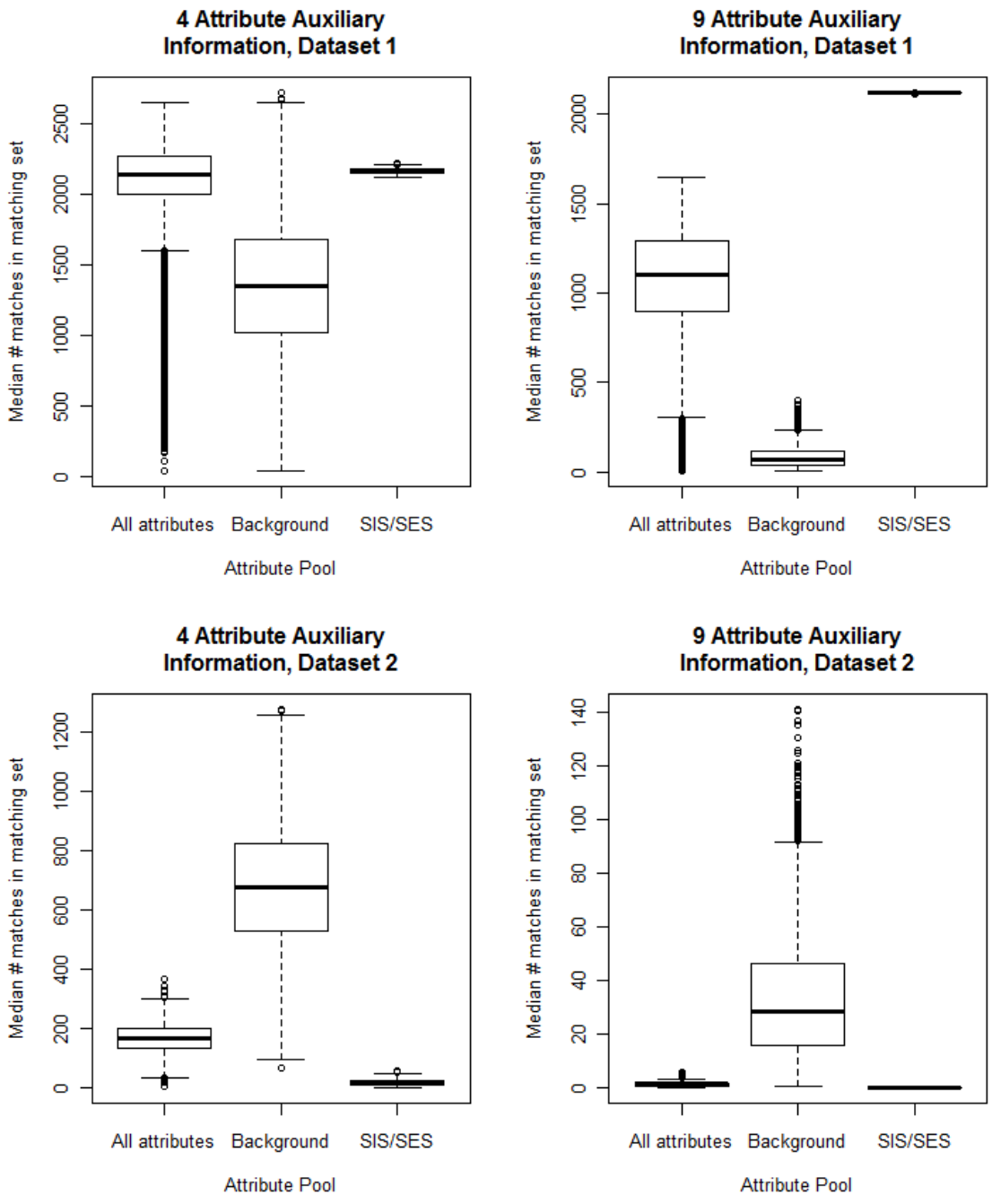

Figure 5: Results of experimental re-identification when the adversary's knowledge is exact. The auxiliary information was drawn randomly from specific attribute pools. The box and whisker plots give the median number of matching records for the test condition and the spread, showing the first and third quartiles and the minimum and maximum matching set size. Points outside the minima and maxima are outliers.

similar overall. Despite higher uniqueness, more than 84% of the participants in Dataset 2 are at least 40% similar overall.

Scaled modules contribute the most to overall similarity, with some participants providing the exact same responses to questions from these modules. Background information (demographics, health information, and sexual background) generates the least similar nearest neighbor scores. Given this information we conclude that questions that are most likely to uniquify participants are also least likely to contribute to similarity.

Our random approach to experimental re-identification selects attributes at random from various attribute subsets.

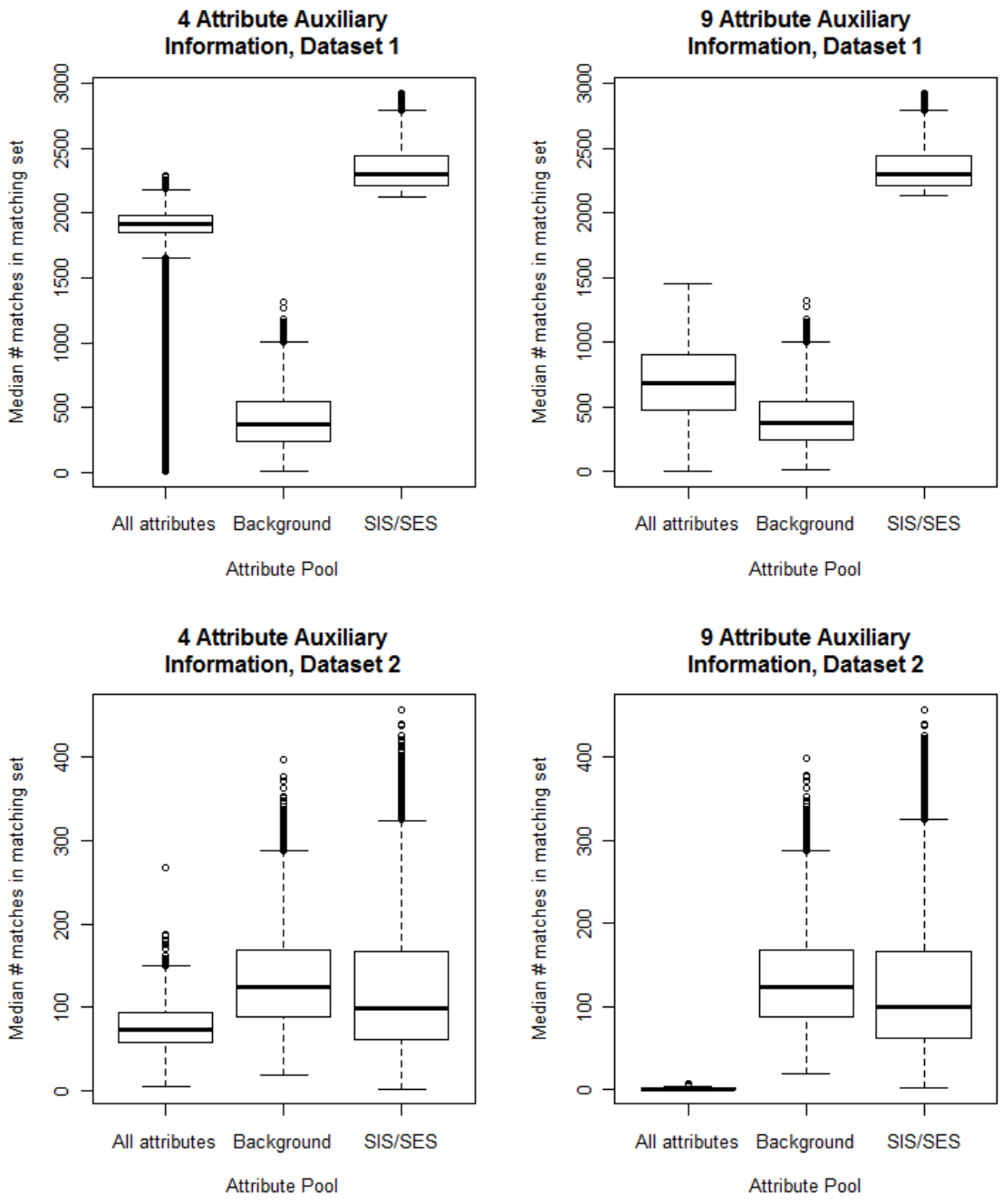

Figure 6: As in Figure 5, but error was added to simulate an adversary's inexact knowledge.

These experiments tend to yield large matching sets, which limits an attacker's ability to re-identify participants. Experimental re-identification using only scaled modules gives much lower success rates than unscaled, as seen in Figures 5 and 6 for Dataset 1. Dataset 2 matching set sizes are smaller than the corresponding sizes in Dataset 1. Therefore we conclude that participants in a sample drawn from more diverse population are exposed to greater privacy risks.

When auxiliary information consists of the most uniquifying attributes and participants are drawn from a diverse population, an attacker is more likely to re-identify participants. Our results show that for Dataset 2, 93% to 100% of the participants can be re-identified. On the other hand, given the same test scenario, no participants from Dataset 1 can be re-identified. Although there are no re-identifications for

Dataset 1, the size of the matching set was usuall very small; a size of 2 was common.

To create a more realistic scenario, we add error to the auxiliary information. Adding error has two interesting side-effects: false positives and misses. False positives occur when the size of the matching set is one, but the record is not the same as the source of the auxiliary information. Misses occur when the size of the matching set is zero, that is, when adding error has brought the auxiliary information out of range of all records. Miss rates could be considered a measure of sparsity. The amount of error we add to variables is small. The fact that this amount of error brings auxiliary information outside the range of all records means that unique values are far apart from each other. Our uniqueness similarity results show that participants tend to be unique in the same attributes, at least for singletons and combinations of two attributes. It appears that this trend does not continue up to four and nine attribute combinations.

5. FUTURE WORK

We have performed extensive analyses of two social science datasets from The Kinsey Institute. We have investigated uniqueness, similarity, and experimental re-identification attacks. Our results show high uniquification rates: nearly 100% of participants have at least one unique three-attribute combination. However, the remainder of their records tends to be very similar. Scaled modules tend to have much lower uniquification rates and background information has the highest.

Overall, most participants' nearest neighbor is more than 50% similar (more than 40% in Dataset 2). In module-specific similarity, scaled modules are predictably far more similar. Experimental re-identification attacks show that there are relatively few successful re-identifi-cations when auxiliary information is randomly chosen. For Dataset 2, re-identification is more likely when auxiliary information is chosen from the most uniquifying attributes. Dataset 1 shows more resistance to re-identification, but the matching set size is still low.

Our goal is to specify a protection model and mechanism for social science datasets. Our results suggest that aggregating data based on its uniqueness properties, whether within modules or the entire dataset, could mitigate privacy risks. We suggest exploring privacy-preserving techniques such as differential privacy [3] [10] and its impact on data utility. We will investigate whether the most uniquifying attributes are available online or in other datasets.

Ultimately, we believe that preserving participant privacy in these datasets and ones like them will be an exercise in risk management and balance. We must decide what is an acceptable level of risk, taking into account both data vulnerability and its required utility in the research for which it was collected.

6. ACKNOWLEDGMENTS

This work is funded by NSF grant CNS-101 2081.

7. REFERENCES

[1] M. Barbaro and T. Zeller. A face is exposed for aol searcher no. 4417749, Aug 9 2006.

[2] L. A. Clark and D. Watson. Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7, 1995.

[3] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin / Heidelberg, 2006.

[4] B. Malin. Re-identification of familial database records. In *AMAI Annual Symposium Proceedings 2006*, pages 524–528, 2006.

[5] B. Malin and L. Sweeney. Re-identification of DNA through an automated linkage process. In *Proceedings of the American Medical Informatics Association Fall Symposium*, pages 423–427, 2001.

[6] B. Malin and L. Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*, 37:179–192, 2004.

[7] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. *IEE Symposium on Security and Privacy*, 0:111–125, 2008.

[8] A. C. Solomon, R. Hill, and E. Janssen. Poster: Privacy and de-identification in high-dimensional social science datasets, 2011. Presented at IEEE Security and Privacy 2011.

[9] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10:557–570, October 2002.

[10] Y. Xiao, L. Xiong, and C. Yuan. Differentially private data release through multidimensional partitioning. In W. Jonker and M. Petkovic, editors, *Secure Data Management*, volume 6358 of *Lecture Notes in Computer Science*, pages 150–168. Springer Berlin / Heidelberg, 2010.

[11] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*, pages 531–540. ACM, 2009.